

Comparing Two Measurement Devices: Review and Extensions to Estimate New Device Variability

Brian J Eastwood, Ph.D.
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285
bje@lilly.com

Abstract

There is much literature available on methods for comparing two measurement systems that are supposed to be equivalent. These methods are briefly reviewed in the context of comparing two in-vitro assays for measuring clotting times where a new method is intended to replace an old method. The basic study is a comparison of results run in both assays. With this study it is possible to determine the relative performance of both assays with respect to bias and variability (standard deviation) using techniques described in Bland and Altman (1981) and Lin (1989) to determine if they are “in agreement”. But it is not possible for such a study to describe the bias or variability of either assay. Usually there is much historical information available for the “old” method. By incorporating that information it is then straightforward to obtain point estimates of the bias and variability of the new assay. Using distributional assumptions or bootstrap estimates one can then obtain confidence interval estimates and conduct hypothesis tests about the absolute and relative performance of the two assays. The performance of various bootstrap and exact-distribution estimates is compared, first in the context of “demonstrating agreement”, and in the context of examining if the new assay has actually improved performance over the old assay.

1. Introduction

As part of the drug discovery process a series of compounds must be evaluated in one or more in-vitro assays to evaluate a compound’s potency against one or more drug targets. Before that process is undertaken the assay is developed and validated for acceptable use and then monitored by testing of controls repeatedly throughout the life of the assay. Often the assay must undergo changes during the drug development process such as a change in technology, transfer to another laboratory or personnel, or a change in key reagents. All of these changes require that the “new” assay be equivalent to the “old” assay so that results for compounds evaluated in the different assays can be compared.

Current standards of practice recommend that a test-retest study be done on a series of compounds covering the potency range using the old and new assays. Then the

same compounds are run again in the new assay to establish the intra-run variability in the new assay (Eastwood, *et al* 2001). In this paper we use the information from the controls run in the old assay to try to eliminate the need for the extra retest in the new assay. Thus with a single run of each assay we can make decisions about the equivalence of the two assays.

2. Data, Models and Estimators

2.1 Data and Distribution Assumptions

The test-retest study data consist of n random pairs (X_t, Y_t) , where X_t and Y_t are the old and new assay results for compound t , $t = 1, \dots, n$. We assume the model $X_t = \mu_t + e_t$ and $Y_t = \mu_t + f_t$, where μ_t is the “true” mean for compound t in that run, and e_t and f_t are intra-run assay measurement errors for the old and new assays. We further assume that $e_t \sim (\mu_e, \sigma_e^2)$ and $f_t \sim (\mu_f, \sigma_f^2)$, and let $M = \mu_e - \mu_f$, and $S = \sigma_f / \sigma_e$. For some, but not all the estimation work ahead we will require that the assay errors be either normally distributed, or at least a known distribution. Where that assumption is required will be identified as needed.

The mean assay errors, μ_e and μ_f , consist of two components each: The long-run over/under estimation of all compounds (bias) and the between-run measurement error. Within a single test-retest study it is not possible to distinguish the two sources. Instead the combined source is tested as a single entity. Moreover, we assume that $M = 0$. Techniques for estimating M and/or testing hypotheses about it are well known, such as the paired T test or the Wilcoxon Signed-Rank test. Therefore, the objective of this paper is to develop methods for estimating S , and/or testing a hypothesis about it.

This is a special case of the problem of comparing two measurement devices for which there is much literature. In general there is considerable disagreement over the suitability of various methods to identify various problems. See Lin (1989), Lin (1991), Atkinson and Nevill (1997), Bland and Altman (1981), Bland and Altman (1983), and Bland and Altman (1995) for reviews and discussions. We adopt and extend the work of Bland and Altman (1981) to this situation, as the model implicit in their approach does not assume that all compounds have the same mean potency.

The historical information about the old assay consists of an estimate, s_e , of the within-run variability, σ_e , which is derived from a sample size sufficiently large that it can be assumed to be a known constant. This information would typically come from controls repeated in each run and any compounds that have been tested more than once.

2.2. Bland-Altman Model

The paper of Bland and Altman (1981) suggests that agreement be assessed by plotting the difference in the two measurements, $d_t = x_t - y_t$, versus the mean, $m_t = (x_t + y_t)/2$. Then if \bar{d} and s_d are the mean and standard deviation of $d_t, t = 1, \dots, n$ then use $\bar{d} \pm 2s_d$ as the limits of agreement. If these limits are sufficiently precise (i.e. tight) then equivalence has been established.

These techniques are based on the assumption that variability is constant over the measurement range, and Bland and Altman suggest that a transformation may be necessary before evaluating the systems. For in-vitro assays transforming the data is usually necessary, and often the log transformation successfully stabilizes the variance. So the Bland-Altman limits are calculated on the log scale and then back-transformed to the linear scale, and the corresponding plot is the Ratio versus the Geometric mean.

2.3. Example

A coagulation assay in current use underwent a technological change. A test-retest study was done on 23 compounds and the potency data is shown in Table 1. Further, a retrospective review of control and re-tested compounds gave a within-run variability estimate $s_e = 0.0434 \mu\text{g/ml}$.

Table 1. Potency Results ($\mu\text{g/ml}$) in a Test-Retest Study on 23 Compounds

t	Old Assay	New Assay	t	Old Assay	New Assay
1	0.175	0.138	13	0.946	0.823
2	0.218	0.177	14	0.925	0.865
3	0.202	0.229	15	0.905	0.904
4	0.178	0.279	16	0.697	1.015
5	0.248	0.294	17	1.142	1.156
6	0.355	0.304	18	1.033	1.224
7	0.430	0.430	19	1.351	1.227
8	0.402	0.478	20	1.530	1.465
9	0.458	0.488	21	1.498	1.650
10	0.515	0.546	22	2.198	2.407
11	0.709	0.615	23	4.337	3.581
12	0.406	0.689			

The mean and standard deviation of the difference in log-potency are 0.0195 and 0.0889 respectively, and a ratio versus geometric mean plot is shown in Figure 1. There is no evidence to suggest that $M \neq 0$, and the limits of agreement are 0.695-1.575. This

is well inside the tolerance limits required by the pharmacologists, which are from 0.33 to 3.0.

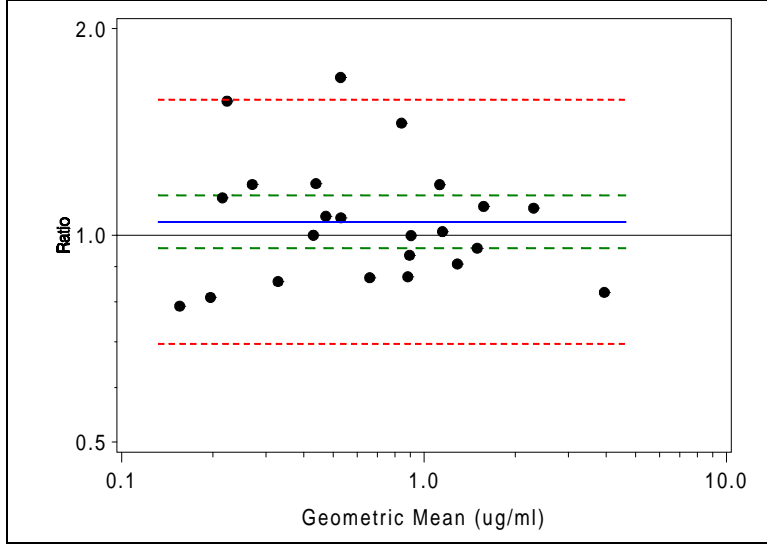


Figure 1. Ratio Versus Geometric Mean Plot for Assay Transfer Data in Table 1. Solid line show mean ratio of new to old. Long dashed lines show 95% confidence interval for mean ratio. Short dashed lines show limits of agreement.

3. Assay Variability Estimators

3.1. Point Estimator

By just using the test-retest study information it is not possible to estimate either σ_e or σ_f as $E(S_d^2) = \sigma_e^2 + \sigma_f^2$. However, using the historical information available to estimate σ_e , we have the resulting estimator

$$s_f = \begin{cases} \sqrt{s_d^2 - s_e^2} & \text{if } s_d^2 > s_e^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This estimator is valid for all assay error distributions with finite variance.

3.1. Normal Theory Interval Estimate

If the assay errors are normally distributed then so are $d_t, t = 1, \dots, n$. Therefore $(n-1)s_d^2 \sim \chi_{(n-1)}^2$. Therefore a normal theory (NT) 95% confidence interval for σ_f is given by

$$CI_{NT} = \left[\max\left(0, \sqrt{CSS_d / \chi_{(n-1, 1-\alpha/2)}^2 - s_e^2}\right), \max\left(0, \sqrt{CSS_d / \chi_{(n-1, \alpha/2)}^2 - s_e^2}\right) \right] \quad (2)$$

3.2. Parametric Bootstrap Estimators

Using a parametric bootstrap (Efron and Tibshirani (1993)) requires making a distributional assumption about the assay errors, but is more general than the normal theory estimator in that one does not have to use the normal distribution to model the assay errors. However, as log-potency values often have approximately normal distributions we will assume normality here.

There are two possible methods for bootstrapping errors. Ultimately the point estimate depends only upon the difference values in log-potency, and therefore one could use a parametric bootstrap sampling distribution for d_t using \bar{d} and s_d as the parameters of the bootstrap distribution. Using these values one can obtain an estimate of the distribution of s_f by generating a number of bootstrap samples. A parametric bootstrap (PB1) 95% confidence interval can be obtained by using the 2.5th and 97.5th percentiles from this distribution. Let CI_{PB1} be this confidence limit.

However, that can result in constant bootstrap estimates if s_d is very small, and therefore a degenerate confidence interval. An alternative parametric bootstrap estimator is to generate bootstrap samples for (x_i, y_i) and then calculate d_t from the generated data. To generate a bootstrap sample for X use $\mu_e = 0$ and $\sigma_e = s_e$. For Y use $\mu_f = \bar{d}$ and $\sigma_f = s_0$, where $s_0 = \max(s_e/5, s_f)$. The lower bound, $s_e/5$, is quite arbitrary. However, σ_f is almost certainly not zero, and highly unlikely in most situations to be more than 5-fold smaller than the old assay variability, and so therefore it seems a reasonable lower bound to impose. Once X and Y samples are generated, estimate \bar{d} and s_d , and then compute the bootstrap sample estimate for σ_f according to (1). Then proceed as above to obtain a parametric bootstrap (PB2) confidence interval, CI_{PB2} . Since the original quantity s_d is not used directly in the bootstrap generator, the confidence interval is much less likely to be degenerate.

3.3. Nonparametric Bootstrap Estimators

A nonparametric bootstrap estimator can be obtained by resampling from the difference values, d_t . Using the percentile method (Davison and Hinkley (1997)) gives a nonparametric bootstrap (NPB) 95% confidence interval, CI_{NPB} . No distributional assumptions are required, however this is an approximate confidence interval.

4. Bootstrap Histograms and Results

We illustrate the bootstrap sample estimators using the data from Table 1 and 3 other datasets. The datasets have decreasing values of s_d , and therefore values of S , as summarized in Table 2.

Table 2. Summary Statistics for the Four Datasets

Statistic	DS 1	DS 2	DS 3	DS 4
n	23	13	19	13
\bar{d}	0.0195	-0.0969	0.0078	-0.0320
s_d	0.0889	0.0839	0.0690	0.0298
s_e	0.0434	0.0788	0.0731	0.0734
\hat{S}	1.79	1.05	0.00	0.00

Histograms of bootstrap sample estimates for each of these datasets are shown in Figure 2(a)-(d) for PB1, 3(a)-(d) for PB2, and 4(a)-4(d) for NPB. In all cases 10000 bootstrap repetitions were used. Note that as \hat{S} decreases there is a greater proportion of samples having a bootstrap estimate of zero. For DS 4, all the bootstrap samples for PB1 and NPB are zero, and hence the bootstrap confidence interval is degenerate in those cases. Also note for datasets DS1 and DS2 the histograms for PB1 and PB2 are somewhat left skewed, whereas the distribution looks symmetric for NPB. This suggests that the underlying distribution may not be normal.

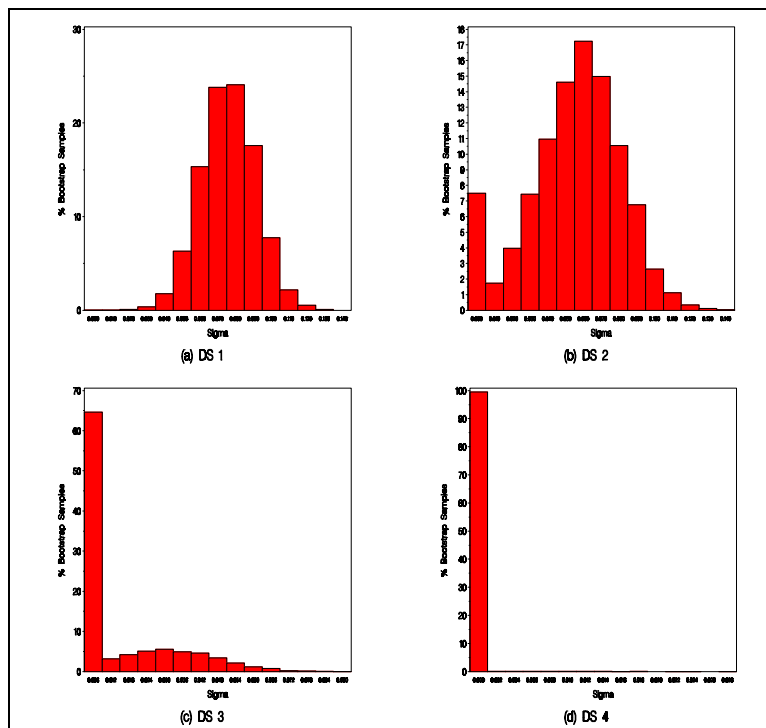


Figure 2. Histograms of PB1 Bootstrap Estimates for Datasets DS1-DS4.

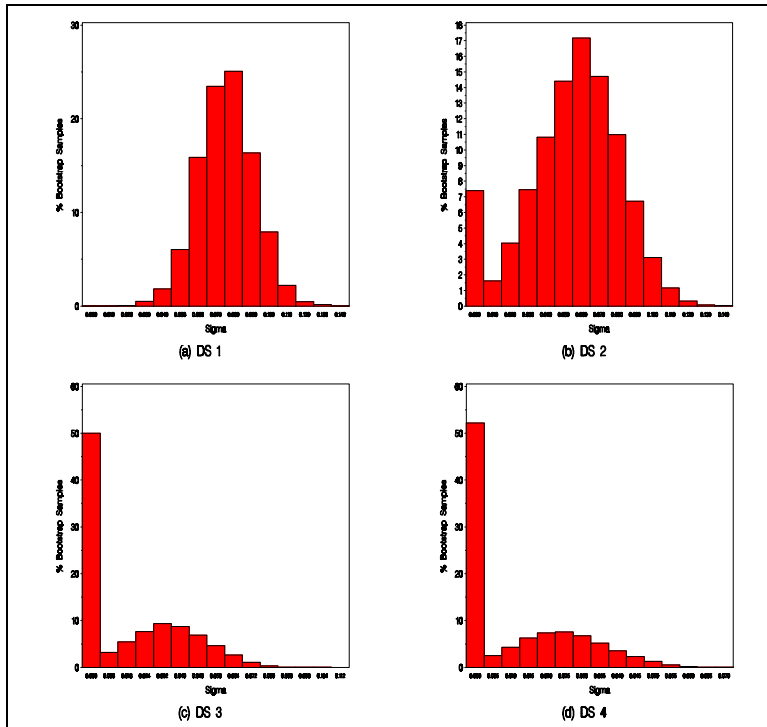


Figure 3. Histograms of PB2 Bootstrap Estimates for Datasets DS1-DS4.

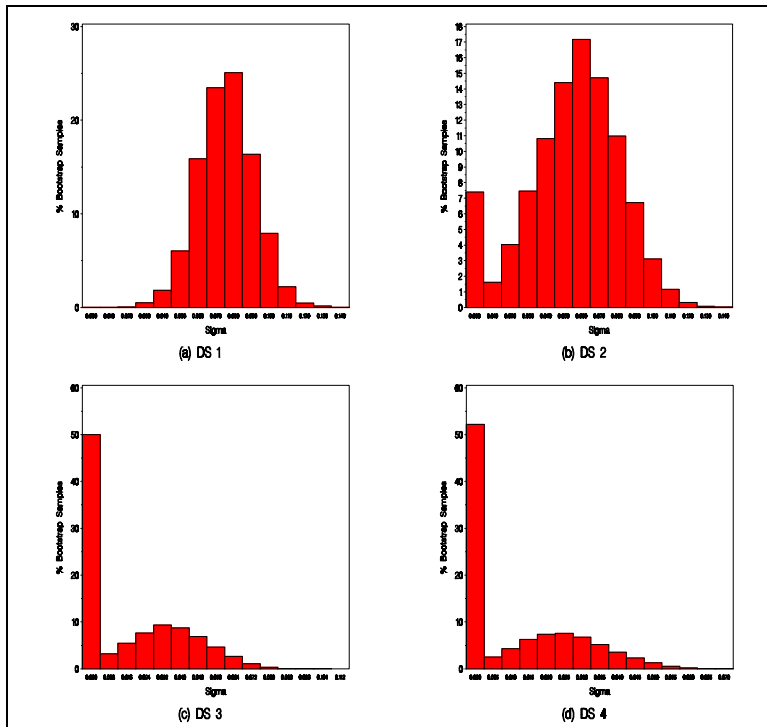


Figure 4. Histograms of NPB Bootstrap Estimates for Datasets DS1-DS4.

Table 3 contains the point estimate and 95% confidence limits for Dataset 1. All suggest that the variability for the assay has increased. Only CI_{NPB} includes the historical estimate (0.0434) within the confidence interval. As indicated above, there is some evidence of non-normality in the assay error distributions, and the Type I error rates may differ from the nominal levels at the sample size used here ($n = 23$). Either or both factors may contribute to the differing results.

Table 3. Point Estimate and 95% Confidence Intervals for σ_f

Quantity	Result
s_f	0.0776
CI_{NT}	(0.0533-0.1181)
CI_{PB1}	(0.0452-0.1060)
CI_{PB2}	(0.0452-0.1060)
CI_{NPB}	(0.0375-0.1019)

CI_{PB1} and CI_{PB2} are equal in this case, although that is not true in general. For example, for Dataset 4 $CI_{PB1} = (0,0)$ and $CI_{PB2} = (0,0.0457)$. The confidence interval may degenerate when the variability in the test-retest study is much smaller than the historical variability of the old assay.

5. Power Study

In order to investigate the power and Type I error properties of the confidence intervals a Monte Carlo power study was conducted for the NT, PB1, PB2 and NPB methods.

A sample size of 23 was used and “true” mean potencies were set equal to the average of the old and new test-retest assay results. In practice the values used are irrelevant for all four methods as all test statistics depend only upon the difference between old and new assay potencies, and are independent of the actual potencies. The true old assay standard deviation, σ_e , was set to s_e , and σ_f was varied from $\sigma_e/5$ to $5\sigma_e$. 10000 bootstrap repetitions were used, and 1000 Monte Carlo repetitions were used. Normally distributed errors were used throughout.

Results are shown in Figure 5. Only the NT procedure achieves the nominal Type I error rate, although BP1 and PB2 come very close (7.5%). The NPB Type I error rate is approximately 15%, too high to be used in practice. The NT is also the most powerful against the alternative $\sigma_f > \sigma_e$, but least powerful against the opposite alternative. NPB achieves the best power against the alternative $\sigma_f < \sigma_e$, although once one normalizes the Type I error rate one suspects that this power advantage would

disappear. PB1 appears only marginally more powerful than PB2 for small values of S , and this power gain over PB2 is due to the greater number of degenerate confidence intervals with that method. In practice PB2 would be preferred over PB1 because many fewer intervals would be degenerate. Overall, PB2 achieved an excellent compromise in power against both alternatives, while achieving an acceptable Type I error rate and thus it would be the best overall method.

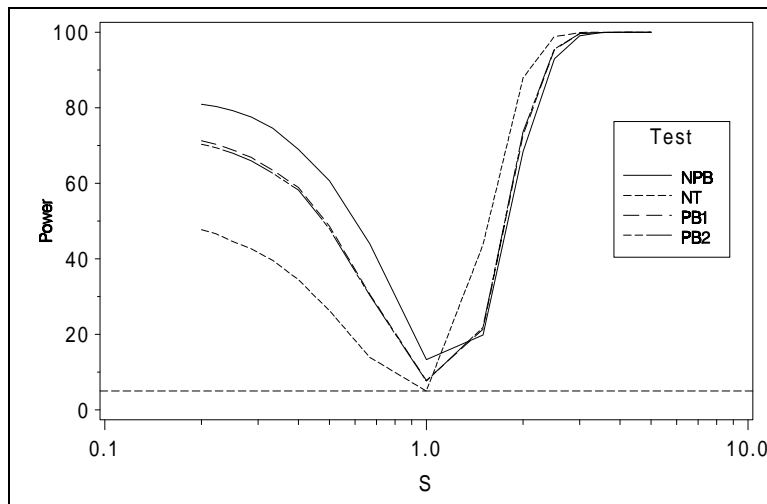


Figure 5. Power of NT, PB2 and NPB test statistics versus S

6. Conclusions

In this paper we examined combining historical data with bootstrap and normal theory confidence intervals. Parametric bootstrap sampling appeared best overall. However, the parametric bootstrap is possibly sensitive to the underlying distribution of assay errors as NT confidence intervals are known to depend upon the assay error distribution and therefore parametric bootstrap sampling intervals presumably do so as well. Further work is needed to investigate this. Note that unlike NT methods which are sensitive to the normal distribution assumption (Armitage and Berry (1994)), parametric bootstrap methods readily generalize to other distributions by replacing the normal random number generator with whatever distribution one desires. Thus using PB2 with appropriate error distributions would appear to be the most acceptable choice. However, Type I error rates need to be examined for alternative error distributions before this conclusion should be accepted.

NPB intervals did not perform well. The high Type I error rate appears to be due to sample size. Further simulations with a larger sample size showed that the Type I error rate decreased as n increased, and was approximately 5% for $n \geq 75$ (data not shown). However, this is too large a sample size to be useful in test-retest studies.

Further work is underway exploring use of better approximation methods to improve the Type I error rate.

Acknowledgements

I would like to thank Gerald F. Smith and Trelia J. Craft at Eli Lilly and Company for supplying the data for the examples, and Wendell C. Smith at Eli Lilly and Company and Raymond J Carroll at Texas A & M University for helpful comments.

References

- Atkinson, Greg and Nevill, Alan (1997), Comment on the Use of Concordance Correlation to Assess the Agreement Between Two Variables, *Biometrics* **53**, 775-778.
- Armitage, P and Berry G (1994), *Statistical Methods in Medical Research*, 3rd Edition, Blackwell Scientific: Oxford U.K., pp. 85-88.
- Bland, J Martin and Altman, Douglas G (1981), Statistical Methods for Assessing Agreement Between Two Methods of Clinical Assessment, *The Lancet* **i**, 307-310.
- Bland, J Martin and Altman, Douglas G (1983), Measurement in Medicine: the Analysis of Method Comparison Studies, *The Statistician* **32**, 307-317.
- Bland, J Martin and Altman, Douglas G (1995), Comparing Two Methods of Clinical Measurement: A Personal History, *International Journal of Epidemiology* **24**, S7-S14.
- Davison, AC and Hinkley, DV (1997), *Bootstrap Methods and Their Application*, Section 5.3.1, Cambridge University Press: Cambridge UK, pp. 202-203.
- Eastwood, Brian J, Farmen Mark W, Iversen, Philip W, Craft, Trelia J, Smallwood, Jeffrey K, and Smith, Gerald F (2001), Methods for the Use and Analysis of Test/Retest Studies to Establish Assay Reproducibility and Equivalence of Two Potency Assays (Poster), Society of Biomolecular Screening Annual Conference, Baltimore MD.
- Efron, Bradley and Tibshirani, Robert J (1993), An Introduction to the Bootstrap, Section 6.5, Chapman and Hall: New York, pp. 53-56.
- Lin, Lawrence I-Kuei (1989), A Concordance Correlation Coefficient to Evaluate Reproducibility, *Biometrics* **45**, 255-268.
- Lin, Lawrence I-Kuei (1992), Assay Validation Using the Concordance Correlation Coefficient, *Biometrics* **48**, 599-604.
- Lin, Lawrence I-Kuei (2000), A Note on the Concordance Correlation Coefficient (Letter), *Biometrics* **56**, 325-325.